FIRE
Forum for Information Retrieval Evaluation (for Indian Languages)

Mandar Mitra
Prasenjit Majumder

Indian Statistical Institute

Kolkata

Overview

- The CLIA project
- Evaluation
  - corpora
  - topics
  - relevance judgments
- Timeline

The CLIA project

- Sponsored by the Dept. of Information Tech., Govt. of India
- Sanctioned in August 2006, work started in early 2007
- Consortium mode project
  - Anna University - College of Engg., Guindy
  - Anna University - KBC centre
  - CDAC - Noida
  - CDAC - Pune
  - **IIIT Hyderabad**
  - **IIT Bombay** (coordinating instt.)
  - **IIT Kharagpur** (co-coordinating instt.)
  - **Indian Statistical Institute**
  - **Jadavpur University**
  - Utkal University

The CLIA project

Assigned task: Create a portal where

1. a user will be able to give a query in one Indian language;
2. s/he will be able to access documents available in the language of the query, Hindi (if the query language is not Hindi), and English,
3. all presented to the user in the language of the query.

Languages

- **Bangla**
- **Hindi**
- Marathi
- Punjabi
- Tamil
- Telugu

Language issues

- Inflectionality:
  Hindi (low) $\rightarrow$ Bangla (medium) $\rightarrow$ Tamil / Telugu (high)
- Spelling variations:
  - case markers may / may not be attached to word
  - long vowels / short vowels, three sibilants, two *N*'s
- Words in a compound may be written together or separately
  e.g. *state government* vs. *StateGovernment*
- Names are often abstract nouns (qualities) / adjectives
  e.g. *Mamata*, *Atal*

Evaluation

Tasks

- Ad-hoc monolingual retrieval for each of the 6 languages
- Ad-hoc cross-lingual retrieval (6×6)

Evaluation

Corpora

- News corpora from 2004-2007 for each language (in UTF-8)
- Plus all available documents on health and tourism
- Bangla corpus:
  - ABP Sep 2004 - July 2007 (186,513 docs., 3.9 GB)
  - CRI Sep 2004 - July 2007 (23,862 files, 124MB)
- *Need to work out distribution issues*

Evaluation

Topics

- Single set of 80 topics (30 training + 50 testing)
- Formulated in English by language representatives
- Translated into all languages
- Deal with national / international issues

Topics

Example:

<title> Political turmoil in South Asian countries

<desc>
Struggle for democratic governance in various South Asian countries.

<narr>
The document should contain information regarding the power struggle between the monarchy / military government and popular political leaders in Nepal, Thailand, Pakistan and Bangladesh.

Topics

Example:

<title> Nobel theft

<desc>
Rabindranath Tagore's Nobel Prize medal was stolen from Santiniketan. The document should contain information about this theft.

<narr>
A relevant document should contain information regarding the missing Nobel Prize Medal that was stolen along with some other artefacts and paintings on 25th March, 2004. Documents containing reports related to investigations by government agencies like CBI / CID are also relevant, as are articles that describe public reaction and expressions of outrage by various political parties.

Topics

- Initial queries formulated by browsing the corpus
- To be refined based on initial retrieval results
- Aim: balance of easy, medium and hard queries

Relevance judgments

Pooling

- Term-weighting + inner product similarity (Smart)
- Language modeling (Lemur)
- Boolean queries
- Divergence from randomness (Terrier)
- Logistic regression (Cheshire)
- Interactive retrieval
- Cover detection (?)

Timeline

29.10.2007	Topic set frozen
31.01.2007	Training data pools complete
16.05.2008	Relevance judgments complete (training data)
01.07.2008	Test topics released
15.09.2008	Runs due (earlier date possible?)
01.12.2008	Results out
15.12.2008	NTCIR
19.12.2008	FIRE?

Acknowledgments

- Noriko Kando - EVIA 2007
- Djoerd Hiemstra, Doug Oard, Mark Sanderson - SIGIR 2007
- Carol Peters - CLEF 2007
- Donna Harman, Ellen Voorhees - TREC 2007
- Google Research Award - support for travel to CLEF 2007